DOCUMENT RESUME

ED 218 920 FL 012 803

AUTHOR Clark, John L. D.
TITLE The Performance of Native Speakers of English on the Test of English as a Foreign Language. TOEFL Research Reports, 1.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE Nov 77
NOTE 40p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Bound Students; Educational Research; *English (Second Language); Higher Education; *Language Tests; *Native Speakers; Psychometrics; Test Construction; *Test Interpretation; Young Adults
IDENTIFIERS *Test of English as a Foreign Language

ABSTRACT

Recent forms of the new three-section Test of English as a Foreign Language (TOEFL) were administered to 88 native speakers of English just prior to graduation from a college-preparatory high school program. Total test score distributions were highly negatively skewed, reinforcing findings of earlier studies that the TOEFL is not psychometrically appropriate for discriminating among native speakers of English with respect to English language competence. Although the native English examinees achieved high total test scores and performed excellently on the listening comprehension section, a number of items in the other two sections (structure and written expression; reading comprehension and vocabulary) were answered incorrectly by over 20 percent of the examinee group. Included in these sections were a number of questions considered by the TOEFL test development staff as dealing with basic grammar points or straightforward reading passages of a type that college-level students should be expected to handle without difficulty. From these results, it is concluded that although response data from native English speaking examinees may be of some use in designating particular test questions for closer examination, errors made by college-bound native speakers should not automatically be considered indicative of item inappropriateness for the TOEFL population. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED218920

TOEFL Research Reports

REPORT 1
NOVEMBER 1977

THE PERFORMANCE OF NATIVE SPEAKERS OF ENGLISH ON THE TEST OF ENGLISH AS A FOREIGN LANGUAGE

John L. D. Clark
Educational Testing Service

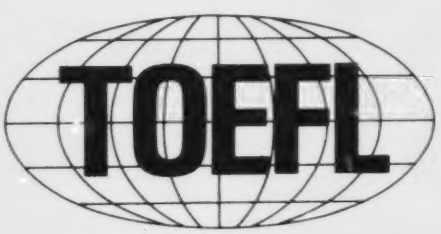
U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Educational Testing Service
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."



CS 012 803

Copyright © 1977 by Educational Testing Service. All rights reserved.

Abstract

Recent forms of the new three-section TOEFL were administered to a total of 88 native speakers of English just prior to graduation from a college-preparatory high school program. Total test score distributions were highly negatively skewed, reinforcing findings of earlier studies that the TOEFL is not psychometrically appropriate for discriminating among native speakers of English with respect to English language competence.

Although the native English examinees achieved high total test scores and performed excellently on the Listening Comprehension section, a number of items in the other two sections (Structure and Written Expression; Reading Comprehension and Vocabulary) were answered incorrectly by over 20 percent of the examinee group. Included in these were a number of questions considered by the TOEFL test development staff to deal with basic grammar points or straightforward reading passages of a type that "college-level" students should be expected to handle without difficulty.

From these results, it is concluded that although response data from native English speaking examinees may be of some use in designating particular test questions for closer examination, errors made by college-bound native speakers should not automatically be considered indicative of item inappropriateness for the TOEFL population.

The Performance of Native Speakers of English on the Test of English as a Foreign Language

John L. D. Clark
Educational Testing Service

A basic principle underlying the development of the Test of English as a Foreign Language (TOEFL) and the interpretation of TOEFL test scores is that the language tasks presented should be those which are relevant to the effective linguistic functioning of non-native English speakers working or studying in the United States. On the assumption that any native speaker of English with a reasonable level of general education would be fully capable of meeting this standard of language performance (and, in most cases, substantially exceeding it), it was considered useful, on an experimental basis, to utilize a native-English speaking population as a measurement "filter" for analyzing the level of difficulty and general appropriateness of test question types currently used in the TOEFL. If native speakers of English at a suitable education level were found to experience virtually no difficulty with the test questions, this would tend to support the assumption that the TOEFL questions are not inappropriately difficult or overly sophisticated for a non-native speaker population. If, on the other hand, native speakers of English were themselves found to have difficulty in responding to particular test sections or question types in the current TOEFL, those portions of the examination would warrant close scrutiny as possibly inappropriate for the measurement purpose intended. If any such difficulties were indeed identified, it was hoped that the patterning of test question responses, together with other types of information on the examinees' performance and approach to the testing task, would provide information that could be used to advantage in modifying and improving the test development process and the content of the examination.

Prior Studies of Native Speaker Performance

An earlier and somewhat related study involving the administration of the TOEFL to native English speakers was conducted by Angoff and Sharon (1971). This study involved the administration of an earlier (1969) form of the TOEFL to 71 entering freshmen at a western state university. Although the mean English achievement level for this group as measured by the American College Testing Program college entrance tests was fairly low by comparison to national ACT norms, the group performed extremely well on the TOEFL from the point of view of total test scores. Notwithstanding a very high overall performance by the examinee group, Angoff and Sharon found that a number of items exhibited "percent-correct" figures of less than 100, indicating that on a number of individual questions at least, the native English speakers were experiencing some difficutly.

Unfortunately, these "problematical" questions were not analyzed in detail in the Angoff-Sharon study, which concentrated on whole-test statistics and normative comparisons between native English speaking and foreign candidate groups. Thus, although the study did confirm the presence of certain questions within the test which were of questionable validity as measures of basic language proficiency, it did not classify or analyze the questions in the detailed manner intended by the present study.

A second study, essentially replicating the Angoff-Sharon study with 173 American freshman and sophomore students at the University of Tennessee, was carried out by Dixon C. Johnson (Johnson, 1977). As in the Angoff-Sharon report, analysis of results was focused on total- and part-score data and did not involve a detailed investigation of individual questions or question types.

Both the Angoff-Sharon and Johnson studies were based on the five-section examination format that was being used in the TOEFL at that time, consisting of Listening Comprehension (50 items), English Structure (40), Vocabulary (40), Reading Comprehension (30), and Writing Ability (40). The present study utilized the new three-section TOEFL described below.

The "New-Format" TOEFL

In early 1975, following studies of the coverage and efficiency of the examination (Pike, 1974), the decision was made to revise somewhat the content and format of the TOEFL, reducing the total number of sections to three and modifying certain of the item types used. The structure of the "new-format" TOEFL is shown below, together with the student directions and example questions for each item type.

Section I: Listening Comprehension (40 minutes, self-paced by tape recording of stimulus materials)

Part A ("Statements") 20 questions

Directions: For each problem in Part A, you will hear a short statement. The statements will be spoken just one time. They will not be written out for you, and you must listen carefully in order to understand what the speaker says.

When you hear a statement, read the four sentences in your test book and decide which one is closest in meaning to the statement you have heard. Then, on your answer sheet, find the number of the problem and mark your answer.

You will hear: John dropped the letter in the mailbox.

You will read: (A) John sent the letter.
(B) John opened the letter.
(C) John lost the letter.
(D) John destroyed the letter.

Sentence (A), "John sent the letter," means most nearly the same as the statement "John dropped the letter in the mailbox." Therefore, you should choose answer (A).

Part B ("Dialogues") 15 questions

Directions: In Part B you will hear fifteen short conversations between two speakers. At the end of each conversation, a third voice will ask a question about what was said. The question will be spoken just one time. After you hear a conversation and the question about it, read the four possible answers and decide which one would be the best answer to the question you have heard. Then, on your answer sheet, find the number of the problem and mark your answer.

You will hear: (man) Hello, Mary. This is Mr. Smith at the office. Is Bill feeling any better today?

(woman) Oh, yes, Mr. Smith. He's feeling much better now. But the doctor says he'll have to stay in bed until Monday.

(third voice) Where is Bill now?

You will read: (A) At the office.
(B) On his way to work.
(C) Home in bed.
(D) Away on vacation.

From the conversation, we know that Bill is sick and will have to remain in bed until Monday. The best answer, then, is (C), "Home in bed." Therefore, you should choose answer (C).

Part C ("Listening Passages") 15 questions

Directions: In this part of the test, you will hear several short talks and conversations. After each talk or conversation, you will be asked some questions. The talks and questions will be spoken just one time. They will not be written out for you, so you will have to listen carefully in order to understand and remember what the speaker says.

When you hear a question, read the four possible answers in your test book and decide which one would be the best answer to the question you have heard. Then, on your answer sheet, find the number of the problem and fill in (blacken) the space that corresponds to the letter of the answer you have chosen.

Listen to this sample talk.

You will hear: People who search for unusual rocks and semiprecious stones are sometimes called rock hounds. Rock hounding is becoming more and more popular as a hobby in the United States. There are over 1,600 rock hound clubs around the nation, and their membership represents only about 7 per cent of the people who participate in the hobby.

The state of New Mexico has devoted a unique state park to the rock hounds. People who visit the park can take samples of rocks away with them. Most parks forbid the removal of any rocks or other materials. Among the rocks found in the New Mexico park are amethysts, opals, and agates.

You will hear: What is the topic of the talk?

You will read: (A) A popular hobby.
(B) The state of New Mexico.
(C) An unusual kind of animal.
(D) New kinds of clubs to join.

The best answer to the question "What is the topic of the talk?" is (A), "A popular hobby." Therefore, you should choose answer (A).

Section II: Structure and Written Expression (25 minutes)

Part A ("Sentence Completion") 15 questions

Directions: In Part A each problem consists of an incomplete sentence. Four words or phrases, marked (A), (B), (C), (D), are given beneath the sentence. You are to choose the one word or phrase that best completes the sentence. Then, on your answer sheet, find the number of the problem and mark your answer.

If John needs a pencil, he can use one _____.
(A) of me
(B) my
(C) mine
(D) of mine

In English, the sentence should read, "If John needs a pencil, he can use one of mine." Therefore you should choose (D).

Part B ("Error Recognition") 25 questions

Directions: Each problem in Part B consists of a sentence in which four words or phrases are underlined. The four underlined parts of the sentence are marked (A), (B), (C), (D). You are to identify the one underlined word or phrase that should be corrected or rewritten. Then, on your answer sheet, find the number of the problem and mark your answer.

At first the old woman seemed unwilling to accept anything
A B
that was offered her by my friends and I.
C D

Answer (D), the underlined pronoun I, would not be accepted in carefully written English; the form me should be used after by. Therefore, the

sentence should read, "At first the old woman seemed unwilling to accept anything that was offered her by my friends and me." To answer the problem correctly, you would choose (D).

Section III: Reading Comprehension and Vocabulary (55 minutes)

Part A ("Vocabulary") 30 questions

Directions: In each sentence of Part A, a word or phrase is underlined. Below each sentence are four other words or phrases. You are to choose the one word or phrase which would best keep the meaning of the original sentence if it were substituted for the underlined word. Look at the example.

The child raced into the house.
(A) walked
(B) crawled
(C) ran
(D) limped

The best answer is (C), because the sentence, "The child ran into the house," is closest in meaning to the original sentence, "The child raced into the house." Therefore you should mark (C).

Part B ("Reading Comprehension") 30 questions

Directions: In Part B, you will be given a variety of reading material (single sentences, paragraphs, advertisements, and the like) followed by questions about the meaning of the material. You are to choose the one best answer, (A), (B), (C), or (D), to each question. Then, on your answer sheet, find the number of the problem and mark your answer. Answer all questions following a passage on the basis of what is stated or implied in that passage.

Read the following sample passage.

The White House, the official home of the President of the United States, was designed by the architect James Hoban, who is said to have been influenced by the design of a palace in Ireland. The building was begun in 1792 and was first occupied by President and Mrs. John Adams in November 1800. The house received its present name when it was painted white after being damaged by fire in 1814.

Example I.

When was the White House first occupied?
(A) 1776
(B) 1792
(C) 1800
(D) 1814

The passage says that the White House was first occupied in November 1800. Therefore you should choose answer (C).

Example II.

The President's house was first painted white when
(A) President and Mrs. Adams requested that it be repainted
(B) it was repaired following a fire
(C) the architect suggested the new color
(D) it was remodeled to look more like an Irish palace

The passage says that the White House was first painted white "after being damaged by fire." Therefore you should choose (B) as the best completion of the sentence.

The two parallel forms of the TOEFL used in this study corresponded to the outline shown above and comprised the initial "base" forms of the new-format examination which were used for norming and scaling purposes in the spring of 1976. As such, they reflect the exact types of items, numbers of items, and sequence of administration used in the operational forms of the new examination.

TEST POPULATION

The native speaker population to which these two forms were administered in the study consisted of graduating American high school students. Participants were drawn from two New Jersey high schools: a Catholic school in a metropolitan center (Trenton) and a regional public school in the immediate suburban area. All of the examinees were students in the senior year, graduating in the spring of 1976. Participation in the project was arranged through the administrative staff at each school, who made announcements of the study and arranged for students to sign up for the testing sessions. Students from the suburban school were offered a payment of $8.00 for participation in the study. At the Catholic school

an equivalent amount was made available for each student tested, to be set aside for use on a "class fund" basis. Although participation in the study was voluntary, observations made by the staff of both schools indicated that the students who agreed to participate were generally representative of the student body as a whole in terms of IQ, courses studied, and high school achievement record.

Even though the principal focus in the study was to be on the performance of college-bound seniors--the native speaker group considered to reflect most closely the level of academic accomplishment and linguistic proficiency at issue in the TOEFL--it was considered advisable for both psychological and public relations reasons to extend the offer of participation to include all graduating seniors. Of the total of 105 students tested across the two schools, 88 indicated on a background questionnaire administered prior to testing that they intended to enter a two- or four-year college in the coming fall, and 17 noted other plans (for example: full-time employment following graduation; entry into a job-related program such as nursing) or indicated that they had not yet formulated their plans. Summary test statistics for the "non-college-bound" group are shown in Table 1 for general comparison purposes but these data are not analyzed further in view of both the very small number of cases involved and the peripheral relevance of this group to the concerns of the present study. The 88 examinees in the major (college-bound) analysis group included 47 males and 41 females, ranging in age from 16 to 18 (mean age: 17.3). All of the examinees were native speakers of English, as determined operationally by responses to the question "Is English your native language (i.e., the language you learned and used as a child?)."

The extent of participants' exposure to foreign languages, either in the classroom setting or in other potential learning situations, was also addressed in the background questionnaire. All but one of the participants reported some school-based (K-12) study of foreign languages, ranging from a single year to a total of nine years, with a mean of 3.6. Exposure to foreign languages at home was only rarely reported: to the question, "Do either of your parents speak a language other than English at home, or do you frequently hear a language other than English (except in school foreign language classes)?," 73 examinees (83.0%) marked "no." The 15 positive responses mentioned occasional conversations overheard between the participants' parents or between their parents and grandparents in Polish, Hungarian, Italian, Spanish, or Dutch. Only two of the participants mentioned that they themselves spoke the foreign language in these situations. With respect to study or travel abroad that would have provided language learning opportunities ("Have you ever studied in a non-English-speaking country or traveled for more than one month in a non-English-speaking country?"), 81 participants (92.1%) reported that they had not engaged in these activities. Of the seven examinees who responded affirmatively, three did not characterize the nature or extent of their travel or study abroad. The other four reported experiences as follows: "two summers in Guatemala, one studying Spanish"; "2 months in Austria and Switzerland"; "Italy - 2 months"; "went to Peru for 6 months when one year old."

On the basis of the native language/foreign language background questions, it was concluded that all of the participants were native speakers of English, with little or no exposure to languages other than English except in the usual school foreign language courses.

With respect to the participants' formal instructional background in English, in addition to the usual elementary school courses, all had taken English courses at the high school level. The mean number of high school English courses reported (including courses currently being taken in the final spring term) was 5.0, with a range of 2-8. About two-thirds of the participants (67.4%) reported that their grade average in the high school English courses was "B." An "A" average was indicated by 18.6% and a "C" average by 14.0%. These responses, both in terms of the number of high school English courses taken and the grade averages obtained, reflect a participant group having a fairly extensive and apparently successful exposure to formal English study, as would be anticipated for a "college-bound" group.

TEST ADMINISTRATION PROCEDURES

Test administration at the two participating schools was carried out on two separate days in mid-May, 1976, approximately one week prior to graduation. At each school, an uninterrupted session of approximately two hours was arranged so that all aspects of the administration (general instructions, testing, and completion of questionnaires) could take place as a unit. The first 15-20 minutes of the session involved a group meeting of all participants, in which the investigator briefly described the project and the procedures to be followed. Participants were told that an "experimental English test" had been developed involving listening comprehension, reading, and other exercises, and that they were being asked to take the test in order to provide data useful for analyzing and improving the test. It was suggested that many of the questions would be quite easy

to answer, but that some of the questions might be more difficult, so that constant attention and application would be necessary. The test was described only as an "English achievement test," and no indication was given that it was to be operationally used with non-native English students. The test booklet cover read simply "Experimental English Test."

Following this introduction, envelopes containing individual copies of the test booklet, answer sheets, the background questionnaire, and a posttesting questionnaire were distributed. The envelopes were pre-arranged in such a way that participants received one of two alternate versions of the test (Form 1 or Form 2) on a statistically random basis. Participants then moved to one of two testing rooms, depending on the test form received; this separation was necessary in order to permit the loudspeaker administration of the Listening Comprehension section of the test, which used different spoken material for the two forms.

In the individual rooms, more detailed testing instructions were given, closely similar to the instructions used operationally in the TOEFL program. However, in addition to being asked to mark the correct answer to each question, participants were urged to "make special note of any individual questions which seem substantially more difficult than the others," or appear to be "tricky, unclear, or which you really have to think about before marking your answer." For any such question, the participant was asked to make a small check mark opposite the question on the answer sheet. Examinees were assured that "there is no expected total number of check marks for you to make, and the number of check marks will not affect your score in any way." Examinees were also informed that there was no penalty for guessing and that every question should be attempted.

The complete test was then administered in the usual order of sections: Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary. The Listening Comprehension section was self-paced by the tape recorded stimulus materials, requiring approximately 40 minutes. For the Structure and Written Expression section and the Reading Comprehension and Vocabulary section, the official time allotments of 25 minutes and 55 minutes respectively were considered unnecessarily generous for native English speakers. For these two sections, the test administrator carefully monitored the pace of responding for the group and attempted to provide a sufficient but not excessive amount of time for completion of the questions. With slight variations in the different testing rooms, the times actually allotted were about 20 minutes and 45 minutes for the two sections, respectively. Later analysis of the test questions showed insignificant drop-out rates for both sections. For both test forms, all examinees answered every item in the Structure and Written Expression section. In the Reading Comprehension and Vocabulary section, there was a very slight drop-out for the last question of Form 1 and the last three questions of Form 2, with 98 percent of the examinees attempting the item in all four cases.

Following test administration, each participant was asked to fill out a short questionnaire asking for his or her judgments of the relative difficulty of the three sections of the test as well as of the difficulty of individual parts (grouped question types) within each section. Spaces were also provided for the participant to write in comments about individual questions or the test as a whole.

The total testing session thus provided four basic types of data for the study: participant background information as previously described, total-test and individual-item results for two forms of the new-format TOEFL as administered to native English speakers, examinee notations of the difficulty or complexity or individual items, and examinee judgments of the relative difficulty of the different sections and item types within the test.

Whole-Test Results

Performance of the native speaker group[1] on the TOEFL was in keeping with the expectation that on the whole the test would be very easy for this population. The mean total test score on Form 1 was 134.42 and on Form 2, 134.91 (maximum possible score on each test: 150). The range of scores on both forms was quite restricted, with standard deviations of 10.19 and 11.44, respectively (Table 1).

The native speaker results are in clear contrast to the performance of the non-native English foreign student group on which these two forms were initially scaled for operational use. Means and standard deviations for the non-native English group were 89.67 and 22.31, respectively, for Form 1, and 88.48 and 21.50 for Form 2, based on N's of 628 for each form.

Score distributions for the native and non-native English groups are shown in Figure 1. The restricted range and highly negative skew of the native speaker distributions are clearly apparent, and are in keeping with

[1]Here and following, unless otherwise indicated, the examinee group consists of the 88 college-bound seniors across the two participating schools.

similar findings by Angoff and Sharon (1971) and by Johnson (197?) based on an earlier TOEFL form. Both of these studies and the present results suggest that any use of the TOEFL to differentiate among native English-speaking students with respect to English language proficiency, at least for college-bound groups, would encounter "ceiling" and range-restriction problems that would render the test virtually useless for this purpose.

Although the whole-test performance of the native speaker group was extremely high by comparison to that of the normal TOEFL candidate population, not every native English speaker achieved a perfect score on the test. Major interest was thus focused on determining which particular sections of the test or types of test items posed some degree of difficulty for native speakers, as reflected in test-section or individual-item statistics.

Results by Test Sections and Parts

Table 2 shows the mean "percent-fail"[2] rates across two test forms for those items comprising the three broad sections of the TOEFL: Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary.

The Listening Comprehension section was appreciably easier than both the Stucture and Written Expression and the Reading Comprehension and Vocabulary sections, as shown by a mean percent-fail rate of 4.4--approximately 7-10 percent lower than for the other two sections. The observed high degree of success in answering the Listening Comprehension questions was considered

[2](Percentage of examinees incorrectly aswering or omitting the item.)

an indication that, for native speakers at least, the simultaneous use in this section of two language modalities (listening and reading) did not in general pose any real difficulty for the examinees. However, as discussed further in the section on individual item results, two of the three Listening Comprehension items defined as "difficult" on the basis of study results (percent-fail levels above 20) contained complex printed answer options which may have confused some of the examinees independently of their comprehension of the spoken material.

The Structure and Written Expression and Reading Comprehension and Vocabulary sections were of virtually equal difficulty as indicated by mean percent-fail figures--14.6 and 12.1, respectively--and clearly more troublesome for the native speaker group than the listening comprehension items.

In addition to an objective measure of difficulty, as reflected in the percent-fail rates, a subjective report of difficulty was provided by the examinees in the form of a rank ordering of the three sections on the posttesting questionnaire. The examinee instructions were as follows:

> How would you compare the <u>difficulty</u> of the three sections? Please rank the three sections according to how much trouble they gave you in answering correctly, how much you had to think about the answers before responding, how unsure you were about the correct answers, etc. Write "1" opposite the section that seemed <u>most difficult</u>, "2" opposite the section that seemed <u>next most difficult</u>, and "3" opposite the section that seemed <u>least difficult</u>. Please make this "1", "2", "3", ranking even if you are not absolutely sure about your choices.

For each section, across the two forms, a "mean difficulty rating" (Table 2) was calculated as the average of the 1, 2, 3 ratings for that section. For reporting purposes, this scale has been reversed, so that higher numbers indicate greater perceived difficulty.

As is consistent with the percent-fail results, the Listening Comprehension section was clearly viewed by the candidates as the least difficult section (mean subjective ranking of 1.21 out of a possible 3.00), and the other two sections were ranked as appreciably more difficult (2.24 and 2.54, respectively).

Percent-fail and subjective difficulty ratings were also obtained for each of the sub-parts (i.e., individual item types) for each section, as shown in Table 3. The difficulty ratings for each sub-part were based on separate posttesting questionnaire sections in which the examinee was asked to "think for the moment only of [name of section]" and to carry out a similar ranking activity for each item type within that section. Printed examples of each type of item were shown in the questionnaire as a memory aid.

Within the Listening Comprehension section, the lowest mean percent-fail is shown for the Listening Passages, followed by Statements and Dialogues. However, all of these figures are within a range of less than 2 percent, and it is thus probably inappropriate to differentiate among the three item types on the basis of observed difficulty. The subjective difficulty ratings for the three item types indicate that the examinees found the Listening Passages considerably more difficult than either the Statements or Dialogues. This probably reflects the more concentrated attention required to listen to the longer stimulus materials, which

examinees could as a consequence have viewed as being "more difficult" than the much shorter stimuli in the other two parts. It should be emphasized, however, that student performance across the entire Listening Comprehension section was quite high, and any difficulties reported with the Listening Passages were not of sufficient magnitude to affect responses to the items in a meaningful way.

Percent-fail results for Sentence Completion and Error Recognition--the two item types making up the Structure and Written Expression section--indicate that Error Recognition items (17.6 percent-fail) were appreciably more difficult for the native speaker group than the items of the Sentence Completion type (9.4). This ranking is not, however, reflected in the subjective difficulty ratings, which are roughly equivalent for the two item types.

Of the two sub-parts of the Reading Comprehension and Vocabulary section, the Vocabulary items (7.9 percent-fail) were appreciably easier overall than the Reading Passages (16.3), and this ranking is also clearly reflected in the difficulty judgments made by the examinees (1.09 and 1.90 out of a possible 2.00).

Considering the mean percent-fail rates for all item types within the test, the three Listening Comprehension item types were less difficult than any of the others, and were all clustered at about 5 percent-fail. Sentence Completion and Vocabulary occupy a middle ground at about 8-10 percent-fail, and Error Recognition and Reading Passages are clearly the two most difficult item types for the examinee group, with mean percent-fail rates of about 16-18. These two item types, especially, deserve close scrutiny

as to the possible factors underlying these results on the part of native English speakers.

Analysis of Individual Items

The third type of analysis--native speaker responses to individual items within each section--was based on the percent-fail index, as well as on the number of times the examinees checked that item on the answer sheet as being unusually difficult or confusing. The total number of check marks made by individual examinees across the complete 150-item test ranged from 0 to 22 on Form 1 and 0 to 17 on Form 2, with means of 6.93 and 7.42 and standard deviations of 6.35 and 4.98, respectively. Although individual tendencies to mark particular items as "difficult" thus varied appreciably among the examinees, the total number of marks for any given item can be considered an indication of the extent to which that item was perceived--on a whole-group-basis--as being "difficult" or "complicated."

For ease of comparison across test forms, the total of check marks for a given item has been expressed in the examples which follow as the percentage of total possible marks for that item (or equivalently, the percentage of examinees who marked that item as "difficult"). For each item, the percent-fail rate is also shown.

Listening Comprehension:

Across both test forms, only three Listening Comprehension items showed a percent-fail level above 20. Of these, two were of the Statements type and one was from the Dialogues part. It is difficult to determine patterns in a reliable manner on the basis of such a small sampling of items and, as previously discussed, the Listening Comprehension section

as a whole was quite free of problematical items. However, it is intriguing to note that both of the Statement items, as shown below,[3] have sets of answer options that are closely similar in form and have a number of common semantic elements which are "scrambled" in various ways across the options. In order to identify the correct answer, the examinee must carefully read each option to discover which combination of elements accurately reflects the situation described in the statement. Since the options are so closely similar, this may be a frustrating and error-inducing task, especially in view of the relatively short time (12 seconds) available to read each of the options and mark the intended answer.

Percentage of Examinees Considering Item "Difficult" (PD)	Percent Fail (PF)
0.0	25.6

[Script] Woman: Having answered all the questions, Tom left the room.

(A) Tom left because all of his questions had been answered.

(B) As he was leaving the room, Tom answered all the questions.

(C) Tom left after all of his questions had been aswered.

(D) Tom left after he had answered all the questions.

[3]Here and elsewhere, the items shown have been altered somewhat from their original form for test security reasons. Care has been taken, however, to preserve those item features relevant to the matters under discussion.

PD: 11.1 PF: 22.2

[Script] Woman: I'm sure that George would rather swim than play golf today.

(A) George would prefer to play golf today.

(B) George likes swimming better than he likes golf.

(C) George likes golf better than he likes swimming.

(D) George would prefer to swim today.

Structure and Written Expression:

In the Structure and Written Expression section, a total of 22 items--4 from the Sentence Completion part and 18 from the Error Recognition part--had percent-fail levels above 20. These 22 items constitute 27.5% of the total number of Structure and Written Expression items appearing in the two forms of the test, a rather large proportion of "problematical" items.

Scrutiny of the individual items revealed three items which could be considered arguable as to the current validity of the grammatical point tested. One involved the "many-much" distinction which is increasingly less rigorously made, especially in the spoken language:

PD: 2.3 PF: 37.2

No other (A) country has lost so much (B) of its citizens though war (C) as has (D) that unfortunate nation.

The second item was based on the "between-among" distinction which may also be considered somewhat tenuous:

PD: 15.6 PF: 40.0

In the early (A) 1950's, Johannsen, in collaboration (B) with Russell Anderson, began arguing (C) for analogies among (D) naturally occuring climatic events and traditional folklore.

The third "arguable" item dealt with the "who-which" contrast, either of which choices might be considered acceptable depending on whether "the fertility goddess Arana" was felt to embody human or non-human qualities:

PD: 2.3 PF 30.2

The tribal chiefs say that (A) the custom dates from the first appearance of the fertility goddess Arana, which (B) appeared on earth (C) as (D) a young girl wearing a headdress of intertwined leaves and berries.

Two items involved informational redundancy, an aspect which might have gone unnoticed by examinees intent on uncovering errors of a more clearly structural nature:

PD: 34.9 PF: 34.9

The new alarm system can guard any entrance (A) by means of an electromagnetic field (B)--an invisible barrier no human can cross (C) without going undetected (D).

PD: 22.2 PF: 28.9

The emperor moved (A) the commercial center from (B) Thikalos to a new city called (C) Massadana, a site now known today (D) as Massada.

There were, however, several other categories of items involving structural aspects with which native English students about to enter college would be presumed to be familiar, but which were missed with appreciable frequency by the examinee group. These are shown below, classified by the grammatical aspect involved:

"You" vs. "one"

PD: 25.6 PF: 30.2

If you give the workers (A) pneumatic tools, one (B) must be sure that they know how to handle (C) the new equipment and that the volume of work (D) is sufficient for continuous employment.

PD: 28.9 PF: 35.6

The more one knows (A) about (B) the subject, the more (C) you understand the importance (D) of handling the matter very carefully.

Parallelism of construction

PD: 17.8 PF: 40.0

The detective wondered (A) whether he should trust the single witness' account of the crime or to search (B) the neighborhood for (C) some (D) other evidence.

PD: 6.7 PF: 42.2

To avoid (A) a ruined project it is necessary to (B) have adequate lighting, the correct type of paint (C), and clean brushes in order to avoid mistakes, prevent waste, and the accomplishment of a (D) proper job.

Pronominal "that"

PD: 11.6 PF: 58.1

Maxwell's approach, like Bellamy (A), was to discount (B) philosophical explanations and insist that objectivism was simply the best way (C), in fact the only way, of saying what he wanted to say (D).

PD: 27.9 PF: 34.9

------- is something on which we can all agree.

(A) That lack of exercise is harmful to one's health
(B) Lack of exercise is harmful to one's health
(C) Not exercising is harmful to one's health
(D) It is harmful to one's health not exercising

PD: 11.1 PF: 31.1

-------, Rhodesia's wealth was founded on tobacco.

(A) Like the American South
(B) As the American South
(C) Like that of the American South
(D) As was the American South

Verb agreement

PD: 23.3 PF: 41.9

After emerging (A) from hibernation, the garter snake is vulnerable (B) to the attacks of predators until their metabolisms (C) are back to normal (D).

PD: 24.4 PF: 35.6

Whether or not (A) Humboldt's understanding of social conflicts are (B) in any way "scientific" remains (C) a troublesome question (D).

PD: 8.9 PF: 33.3

While remaining accountable (A) to his or her company as well as (B) to those individuals who submit (C) damage claims, an insurance adjuster must decide each case on their (D) own merits.

PD: 4.7 PF: 23.3

That (A) ethnic minority groups and women should be treated (B) equally in government-related matters are now (C) a legal requirement (D).

It is difficult to consider the preceding 11 items as being in any way faulty or misleading, and the suggestion may be advanced that certain areas of deficiency in the grammatical preparation of the examinees for college-level work have been identified, rather than item flaws requiring attention in the test development process.

In the Reading Comprehension and Vocabulary section, 8 Vocabulary items and 14 Reading Comprehension items had percent-fail rates of over 20. The lexical items tested in each of these questions are shown below in order of increasing difficulty for the native English examinee group. For each, the frequency per million on the Thorndike-Lorge (1944) general word list is also shown.

	Difficulty (percent-fail)	T-L Frequency
improved	24	4
incessant	28	5
scope	29	11
broach	31	3
discreetly	33	2
remuneration	35	1
lethargy	37	1
razed	42	2

These results may be compared to the Thorndike-Lorge frequencies for those items which all of the examinees answered correctly:

	Difficulty (percent-fail)	T-L Frequency
lagging	0	9
hurdles	0	2
grinned	0	21
mandatory	0	1
discarded	0	10
hilarious	0	2
coaxed	0	10

Although the first listing would appear to suggest some degree of positive relationship between the lexical rarity of a given vocabulary item (as measured by Thorndike-Lorge) and the difficulty of the item for

the native English examinees, the strength of this relationship may be questioned by the students' unfailingly correct responses on such items as "mandatory," "hurdles," and "hilarious," which are in the two lowest frequency categories of the Thorndike-Lorge list.

With respect to test development implications of these results, and pending more extensive and detailed investigation of the frequency/difficulty relationship for vocabulary items, it may be suggested that a preliminary trial of draft Vocabulary items with a group of native English speakers generally similar to those used in the study could "flag" certain items which are not uniformly answered correctly by the native examinee group. Words so flagged could be further examined by test development staff as to the advisability of including them in operational test forms. On the assumption that items such as "discreetly" and "remuneration" would be fairly useful and productive in real-life contexts typical of those encountered by foreign students in the United States, such items might justifiably by included even though native English speaker performance on the same item was not uniformly perfect. On the other hand, items which are both found to pose some difficulty for native English speakers and judged to be of relatively little utility in a foreign student/resident context (for example, "lethargy," "razed") could be identified and rejected through this review process. Use of the Thorndike-Lorge or other frequency counts might be useful as a general guide to the extent-of-use question, but the judgment of test developers and reviewers highly familiar with language-use requirements in a foreign student setting would justifiably take precedence over raw frequency data.

In the Reading Comprehension section, among the 14 items with percent-fail rates above 20, 5 were found to require a passage summarization or interpretation activity on the part of the examinee, as indicated by item stems containing such phrases as "It can be inferred that...," "Critics... are most likely to accuse it of being...," "The author assumes that...," "The author would be most likely to agree with...," and "The aspect...that had the LEAST continuity was most likely the...." Percent-fail figures for these items ranged from 32.6 to 64.4, and the percentage of students marking the item "difficult" ranged from 11.6 to 22.2.

To investigate this apparent trend in more detail, all of the Reading Comprehension items across the two test forms were reviewed and categorized by the investigator (without reference to the statistical results) as to whether or not they required summarization or interpretation on the examinee's part. On the basis of this count, 50 percent of the total of 10 items identified as involving "summarization/interpretation" were found to have percent-fail levels in excess of 20 percent, and only 18 percent of the remaining 50 items (that is, the other-than-summarization/interpretation items) were found to be "difficult" by the same criterion. Percent-fail rates for the latter category of items ranged from 23 to 53. These results suggest that although "difficult" items in the Reading Comprehension section were predominantly of the type requiring summarization or interpretation on the student's part, a number of the other more factually-oriented questions were also missed with some frequency by the native speaker examinee group.

Summary and Conclusions

Two forms of the "new-format" TOEFL were administered to a population of native-English speaking high school seniors who were preparing to begin undergraduate study the following fall. Whole-test performance of this group on both tests was very much higher than that of the non-native speaker populations with which the test is used operationally. The highly negative skew and resulting "ceiling" effects of the native speaker score distributions obtained in this study reinforce the conclusion that the TOEFL is not an appropriate means of differentiating among educated native speakers of English with respect to native language proficiency.

Of the three major sections of the TOEFL (Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary), Listening Comprehension was easiest for the native English speaker group, with only three of 100 items showing a percent-fail rate higher than 20. Within the Listening Comprehension section, the Listening Passages were perceived by the examinees as being appreciably more difficult than the Statements or Dialogues items, but their actual performance on the passages was not markedly lower than on the other two types of Listening Comprehension items.

Within the Structure and Written Expression section, the Error Recognition item type, in which the examinee reads complete sentences and identifies grammatically incorrect underlined portions, was appreciably more difficult than the Sentence Completion type. Over one-third (36.0%) of the Error Recognition items had percent-fail rates of above 20, raising the question of whether an appreciable number of these items could be "faulty." Analysis of individual items within this and the Sentence

Completion type indicated that a few of the items could be considered to deal with a questionable point of grammar, but most of the items missed by the participants involved quite basic points of grammar, such as parallelism of construction and verb agreement.

In the Reading Comprehension and Vocabulary section, the discrete Vocabulary items were appreciably easier overall than the Reading Comprehension passages (mean percent-fail levels of 7.9 and 16.3, respectively). Although several of the lexical items missed by the native English speakers showed low (1-2 per million) frequencies on the Thorndike-Lorge general list, other words of similar frequency were correctly responded to by the entire examinee group. On the basis of these results, it is suggested that the practical utility of a given lexical item for non-native English student/resident groups should continue to be a major factor in the selection of vocabulary items, and that both native-speaker performance and frequency data should play a secondary role in this regard.

For the Reading Comprehension passages and associated items, percent-fail levels for the native English speakers were as high as 64.4, suggesting substantial difficulty on the part of the examinees in responding correctly to certain of the reading questions. Closer analysis indicated that a high proportion of these items dealt with the summarization or interpretation of passage content, although several of the more factually oriented questions also proved difficult for the examinee group.

With respect to the test development implications of the study results, it can be suggested as a first point that the Listening Comprehension items--including the Statement, Dialogue, and Listening Passage item types as currently used in the TOEFL--present virtually no difficulty for native

English speakers at the college-entry level. By this light, the linguistic content of the spoken materials presented to foreign candidates in the TOEFL is certainly not beyond that which native speakers would be easily able to handle, and these results tend to alleviate any concern that TOEFL candidates are being presented with unfair (i.e., beyond-the-average-native speaker) listening tasks in this portion of the test. It should be noted, however, that the present study does not address the possible existence of a "reading load" as a complicating factor in non-native candidates' responses to listening comprehension material. Although non-native candidates with a fairly high level of English reading proficiency should have little difficulty in reading the printed answer options for the Listening Comprehension section, it is at least theoretically possible that candidates with an appreciably lower level of reading ability might not be able to read each of the answer options with full comprehension in the time allotted, and thus would obtain an inappropriately low listening comprehension score even though the spoken stimuli themselves had been well understood. A study aimed at determining the presence or absence of a "reading load" in this part of the test would be desirable as a follow-up to the present study.

Native English speaker responses to the items in the Structure and Written Expression section were more difficult to interpret in that a number of items which the investigator and other TOEFL staff considered both "basic" and "easy" were missed by a relatively high proportion of the native speaker group. If the performance of this group can be considered representative of college-bound native speakers generally, and if the TOEFL program were to adopt the procedure of eliminating from the examination any

and all items found difficult for this population, it appears that such a procedure would exclude a number of aspects of basic English structure that are at least subjectively indispensable for effective academic work at the undergraduate level.

Study results for the Reading Comprehension and Vocabulary section suggest that certain of the vocabulary items appearing in representative forms of the examination, as well as some of the passage-based reading comprehension questions, pose some difficulty for native English speakers at the college-entry level. While it may be useful for test assembly purposes to "flag" for closer scrutiny items answered incorrectly by a specified percentage of native English speakers, the informed judgment of test developers and reviewers closely familiar with reading materials and other language requirements that will be faced in the U.S. by TOEFL candidates should continue to play an important role in this regard.

Figure 1

Score Distributions of Native English Speaking and Non-Native English Speaking Candidates on Two Forms of the TOEFL

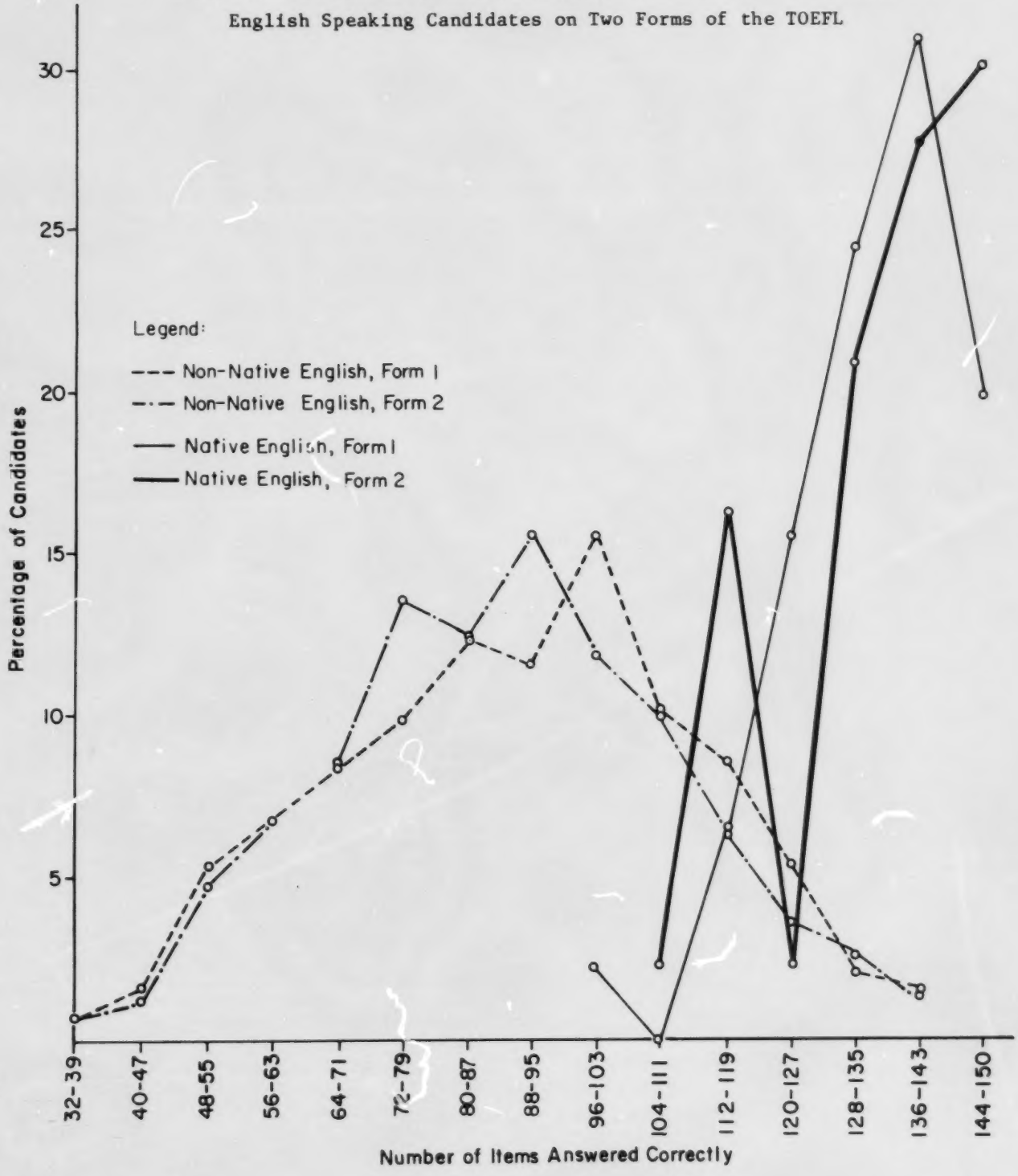


Table 1

Statistical Summary, TOEFL Forms 1 and 2

	College-bound Seniors	Non-College bound Seniors	Combined Group
Form 1			
Number of Cases	45	8	53
Total Items	150	150	150
Obtained Range	103-150	86-145	86-150
Mean Total Score	134.4	121.4	132.5
S.D.	10.2	18.1	12.6
Reliability ($K\text{-}R_{20}$)	.89	.95	.92
Form 2			
Number of Cases	43	9	52
Total Items	150	150	150
Obtained Range	111-150	110-146	110-150
Mean Total Score	134.9	130.2	134.1
S.D.	11.4	12.9	11.8
Reliability ($K\text{-}R_{20}$)	.92	.94	.92

Table 2

Observed and Subjectively Rated Difficulty of TOEFL Sections

for Native Speaker Group

		No. of Items Presented (Forms 1 and 2 combined)	Mean Percent-Fail	Mean Difficulty Rating[a]
I.	Listening Comprehension	100	4.4	1.21
II.	Structure and Written Expression	80	14.6	2.24
III.	Reading Comprehension	120	12.1	2.54

[a] On a scale in which 1 means least difficult, 2 intermediate, and 3 most difficult. (See text for detailed description.)

Table 3

Observed and Subjectively Rated Difficulty of TOEFL Item Types

for Native Speaker Group

	No. of Items Presented (Forms 1 and 2 combined)	Mean Percent-Fail	Mean Difficulty Rating[a]
I. Listening Comprehension			
A. Statements	40	4.7	1.60
B. Dialogues	30	3.3	1.77
C. Listening Passages	30	5.2	2.61
II. Structure and Written Expression			
A. Sentence Completion	30	9.4	1.14
B. Error Recognition	50	17.6	1.09
III. Reading Comprehension and Vocabulary			
A. Vocabulary	60	7.9	1.09
B. Reading Passages	60	16.3	1.90

[a] Possible ranges:

Listening Comprehension: 1-3; Structure and Written Expression: 1-2; Reading Comprehension and Vocabulary: 1-2. For each section, higher numbers indicate relatively greater difficulty.

REFERENCES

Angoff, William H. and Sharon, Amiel T. A comparison of scores earned on the Test of English as a Foreign Language by native American college students and foreign applicants to U.S. colleges. TESOL Quarterly, 5(2), 1971, 129-136.

Johnson, Dixon C. The TOEFL and domestic students: conclusively inappropriate. TESOL Quarterly, 11(1), 1977, 79-86.

Pike, Lewis W. An evaluation of alternative item formats for testing English as a foreign language. Project Report, Educational Testing Service, Princeton, N.J., 1974.

Thorndike, Edward L. and Lorge, Irving. The teacher's word book of 30,000 words. Bureau of Publications, Teachers College, Columbia University, New York, 1944.